COMPARATIVE STUDY OF DATA WAREHOUSE DESIGN APPROACHES: A SURVEY

Rajni Jindal[1] and Shweta Taneja[2]

[1] Associate Professor, Dept. of Computer Engineering, Delhi Technological University
Formerly Delhi College of Engineering (DCE), Bawana Road, Delhi-42.
rajnijindal@dce.ac.in

[2] Research Scholar, Dept. of Computer Engineering, Delhi Technological University
Formerly Delhi College of Engineering (DCE), Bawana Road, Delhi-42.
shweta_taneja08@yahoo.co.in

ABSTRACT

The process of developing a data warehouse starts with identifying and gathering requirements, designing the dimensional model followed by testing and maintenance. The design phase is the most important activity in the successful building of a data warehouse.

In this paper, we surveyed and evaluated the literature related to the various data warehouse design approaches on the basis of design criteria and propose a generalized object oriented conceptual design framework based on UML that meets all types of user needs.

KEYWORDS

Data warehouse design, Multidimensional modelling, Unified Modelling Language

1. INTRODUCTION

A Data Warehouse (DW) is defined as "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process" [1]. Data warehouses store huge amount of information from multiple data sources which is used for query and analysis. Therefore, the data is stored in the multidimensional (M D) structure [34]. A multidimensional model stores information into facts and dimensions. A fact contains the interesting concepts or measures (fact attributes) of a business process (sales, deliveries, etc.), whereas a dimension represents the perspective or view for analyzing a fact (product, customer, time, etc.) using hierarchically organized dimension attributes. Multidimensional modelling requires specialized design techniques that resemble the traditional database design methods [5] as shown in table 1.The database design consists of following five phases. The first phase is Analysis of operational systems whose aim is to collect the information concerning the pre existing operational system. It involves the designer, along with the people involved in managing the information system and produces in output the (conceptual or logical) schemes of either the whole or part of the information system. The next phase consists in gathering and filtering the user requirements. It involves the designer and users of the DW, and produces in output the

DOI: 10.5121/ijdms.2012.4104

specifications concerning the choice of facts and dimensions etc.Next is the Conceptual design [36] which aims at producing an implementation-independent and expressive conceptual schema for the DW, according to the chosen conceptual model. Logical design takes as input the conceptual schema and creates a corresponding logical schema (which is more detailed) on the chosen platform by considering some set of constraints. And finally, a phase of physical design, that takes into account issues specifically related to the tools chosen for implementation –such as indexing and allocation.

Table 1. Database Design Methods

Step	Input	Output
Analysis of Operational System	Information regarding the operational systems	Database schemes
Requirements Elicitation	Database scheme	Specifications for data warehouse
Conceptual design	Database scheme and Specifications	Conceptual schema
Logical design	Conceptual schema	Logical schema
Physical design	Logical schema	Physical schema

Conceptual modelling is the necessary foundation for building a database that is well-documented and fully satisfies the user requirements. In this phase, the multidimensional schema of the data warehouse is defined. It is the beginning of the design process on which the success of entire data warehouse depends. Though several conceptual models have been proposed, none of them has been accepted as a standard so far, and all vendors propose their own proprietary design methods .Obviously, there is a need of existence of a standard conceptual model.

In this paper, we make a comparative study of various approaches and techniques for data warehouse design and propose an Object Oriented framework for the conceptual design of a data warehouse. We have used UML ([2], [3], [4]) in the design process as it has become a standard for object modelling during analysis and design steps of software system development. So, it reduces the effort of learning new notations or methodologies for every subsystem to be modelled.

2. BACKGROUND WORK OF DW DESIGN APPROACHES

In the literature, different data models [24, 25,26] both conceptual and logical have been proposed for data warehouse design .These approaches are based on their own visual modelling languages or make use of well known graphical notation like ER model or UML, but to the best of our knowledge, there is no standard method or model that allows us to model all aspects of a DW. Moreover, during our survey we noticed that most of the research efforts in designing and modelling DWs have been focused on the development of MD data models and conceptual design, the interest on the physical design of DWs has been very poor.

The pioneer author in the field of data warehouse design is Juan Trujilio.He has made a major contribution. He proposed the use of UML for the design of data warehouse. He defined four UML profiles for modelling different aspects of data warehouse: the UML profile for Multidimensional Modelling, the Data Mapping profile, the ETL profile and Database Deployment profile. In [25],authors propose an approach that provides a theoretical foundation for the use of OO databases and Object relational databases in DW.This approach introduces a set

of minimal constraints and extensions to UML for representing multidimensional modelling properties for DW.In [10,27], authors have proposed a multidimensional profile for the Data warehouse conceptual schema and Client conceptual schema. The author has also shown work in the field of physical schema. He has presented the database deployment profile in [13].

Another author who also has a significant role in the design of data warehouse is Stefano Rizzi. The author in [8] proposed a graphical conceptual model for data warehouses, called Dimensional Fact model, and gave a semi-automated methodology to build it from the pre-existing (conceptual or logical)schemes. Then in [11] based on the Dimensional Fact Model (DFM), he gave a general methodological framework for data warehouse design. Then he discussed some issues in Multidimensional modelling for the design of data ware house in [16].After that different authors gave different techniques and models for the design of data warehouse which we have discussed and compared in the next section.

3. COMPARITIVE ANALYSIS OF DW DESIGN TECHNIQUES:

We have analyzed research work done in data warehouse design and its related issues. A brief tabular comparison has been provided below in table 1 on the basis of following criteria: Proposal, Framework/Architecture, Approach or technique proposed, schema used, whether the design can be extended to logical and physical design also, case study and tool used.

Table 1. Comparison of Work done by different Authors

Features → / Authors ↓	Proposal	Framework/Architecture	Approach/Technique	Schema used	Extended to Logical Level	Extended to Physical level	Case Study	Tool Support
1.S.Rizzi, Golfareli (1998) [11]		They propose a graphical conceptual model for DW called Dimensional Fact model	Gave a methodology to build it from the pre existing schemes.	Star	Yes	Yes	Sales DW	---
2. S.Rizzi, Golfareli (1999) [18]	They propose a general methodological framework for data warehouse design, based on Dimensional Fact Model (DFM).	-----	After analyzing the existing information system and collecting the user requirements, a conceptual design is	Star	Yes	Yes	-----	---

			carried out semi-automatically starting from the operational database scheme and then a workload is prepared.					
3.Juan Tujilio,Gomez (2000)	Object oriented approach to accomplish the conceptual modelling Of DW ,MD databases and OLAP application	----	They introduced a set of constraints and extensions to UML for showing MD modelling properties	Star	No	No	Sales DW	UML
4.Golfarelli, M., & Rizzi, S. (2001) [28]	They illustrated the main features of WAND, a prototype CASE tool for data warehouse design.	----	Implemented the tool WAND, which helps in structuring a data mart and carrying out the Conceptual design in a semi-automatic fashion.	----	Yes	-----	----	----
5. Juan Trujilio,E.Medina and S.Lujan Mora [2002] [9]	They show how to manage the representation, manipulation and presentation of MD models on the	-----	They use Object-Oriented (OO) approach based on the	----	----	---	---	CASE tool

	web by means of extensible Style sheet Language Transformations (XSLT).		Unified Modelling Language.					
6.Abello, Saltor (2002) [6]	They present a MD conceptual object oriented model using extension of UML	-----	Used different structures of object oriented model like nodes,arcs ,detailed levels, stereotype s	----	---	----	---	----
7.Juan Trujilio , S.Lujan Mora and I.Song [2002] [10]	They present the development of multidimensiona l (MD) models for data warehouses using UML package diagrams	---	They present design guidelines and explain them with various examples.	---	----	----	---	Ration al Rose 2000
8.Lujan Mora and Juan Trujilio (2003) [12]	They present a DW design method based on UML which is used to handle all DW design phases and steps from operational data sources to final implementation	Object oriented method based on UML	Used MD modelling ,MD databases and OLAP support	---	Yes	Yes	----	----
9. Lujan Mora and Juan Trujilio (2004)[13]	They proposed to model the physical design of a DW by using the component diagram and deployment diagram of UML	Framewo rk of a DW with 5 stages(i.e . source,int egration, DW,cust omisatio n and	Their approach reduces the overall developm ent time of a DW and covers all main	----	Yes	Yes	----	UML

		client) and 3 levels(conceptual, logical and physical). Each level is composed of diagrams -making a total of 15 diagrams	design phases of DW from conceptual modelling till final implementation					
10. Lujan Mora and Juan Trujilio,Vassiladis (2004) [14]	----	Framework for the design of DW back-stage including transformation rules at the attribute level and modelling of relationship between sources and targets in different levels of granularity	Their approach is based on usage of UML packages	---	Yes	Yes	----	Extend UML
11.Rizzi,Trujilio,Abello (2006) [16]	They aimed at discussing some open issues in modelling and design of data warehouses. Like issues regarding conceptual models, logical models, methods for design, interoperability etc.	----	-----	----	---	----	---	---

12.Medina,Trujilio et al.(2006) [31]	They gave a UML profile to represent MD and security aspects of conceptual modelling	-----	Use UML packages to group classes into higher level units	Star	----	----	Health system	UML
13.Deepti Mishra,Ali Yazici,Beril Pinar (2008) [17]	To compare various conceptual and logical DW design models and to find which is more suitable for implementing DW	----	-----	----	----	----	Sales and shipping system	----
14.Kamal Alaskar and Akhtar Shaikh (2009) [18]	Gave a UML profile for modelling DW usage on a conceptual level	Designed a conceptual UML model and translated it into XML logical model which is later converted into XML document as physical model	-----	Star	Yes	Yes	Hajj Pilgrim private tour	UML
15.Hui Ma,Yiping Yang and Fan Zhang (2009) [19]	They gave Anti-standardization technique in DW design	Converted ER model to multidimensional data model	Gave different anti standardization methods like increasing data redundancy, increasing derived columns etc.	Star	-----	----	Stock exchange	----
16.Fernandez Medina et al. (2010) [20]	Proposed a conceptual Multidimensional Data model(called as PIM)and transformed it into a secure XML DW as a logical model(called as PSM)	Model Driven Architecture(MDA) for developing secure DW.	Set of transformation rules to convert conceptual data model to logical model	---	yes	no	Airport DW	XML

17.Payal Pahwa and Shweta Taneja (2010) [23]	UML multidimensional model from various data sources based on UML schema	Conceptual level integration framework based on UML sources. First convert UML schema to UML class diagrams and then build multidimensional model from it	Object oriented approach for DW design. Mapping rules to convert UML class diagram to multidimensional model	Snowflake	----	---	Diabetic Interactive Electronic Treatment System(DIET)	UML
18.Francois Pinet et. al. (2010) [21]	Used UML to build a DW model	-----	Gave an overview on UML based techniques and tools used in agricultural DW	-----	Yes	Yes	Pesticides in agriculture	UML
19.Anjana Gosain,Suman Mann (2010) [30]	-----	They extended the work of Alexander [7].They introduced seven operators over DW model.	Defined object oriented Md data model for description of data	----	No	No	----	UML
20.Jesus Pardillo and Jose Noberto Mazon (2011) [22]	They proposed that the use of ontologies will improve several aspects of the design of data Warehouses.	described several shortcomings of current data warehouse design approaches and discussed the benefit of using ontologies to overcome them	------	----	No	No	---	---

4. PROPOSED OBJECT ORIENTED FRAMEWORK FOR DATAWAREHOUSE CONCEPTUAL DESIGN

4.1 Components of the Framework

Our framework takes into account the requirements of the users. The framework is divided into two levels namely- Requirements level and Design level. At the Requirement level, the requirements are gathered from different users and a thorough analysis is made. The Integrator component integrates the collected requirements. Each level comprises of a number of components to control particular tasks along with detailed metadata repository to speed up the whole process.

In the next level, that is the design level; UML designer helps in extracting major objects and classes from data gathered from multiple data sources and constructs UML class diagrams. The UML class diagrams are converted to multi-dimensional model represented in the form of star or snowflake schema. The conversion is done by applying certain mapping rules that help in mapping the classes to facts and dimensions.

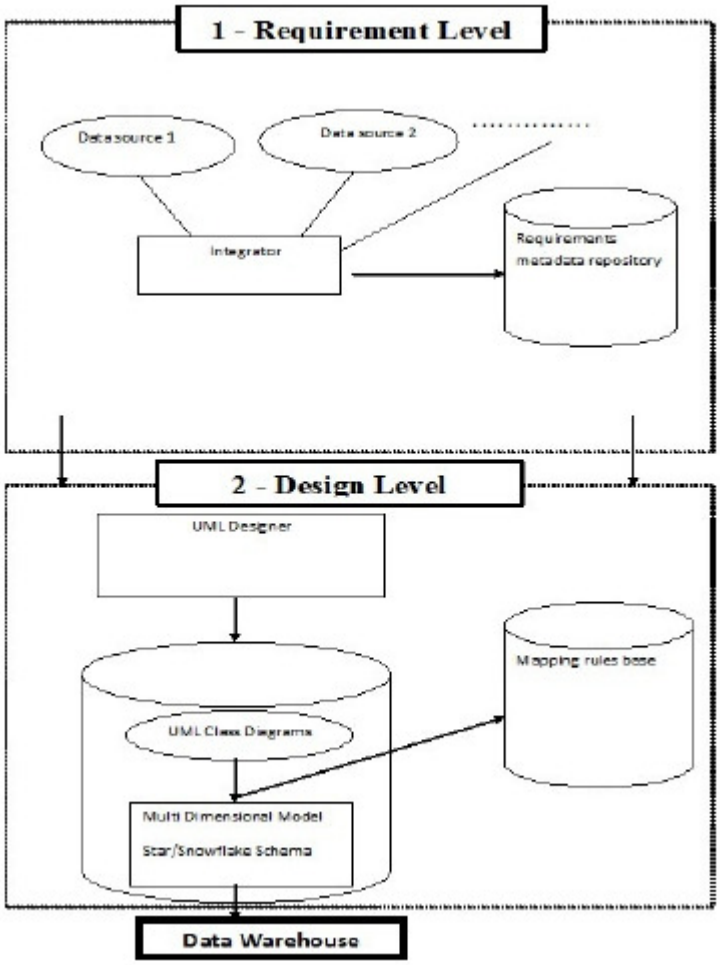

Figure 1. Proposed Framework

5. CONCLUSION AND FUTURE SCOPE

In this paper, we make a comparative study of different approaches used for data warehouse design. In the literature survey, different authors [4, 5, 6, 7,8, 9,10,11,12,13] have proposed different techniques at different levels i.e. conceptual level ,logical level and physical level. Our comparative study is based on following criteria: Proposal, Framework/Architecture, Approach or technique proposed, Schema used, whether the design can be extended to logical and physical design also, Case study and Tool used.

We also propose an Object Oriented framework for data warehouse conceptual design. Our framework has many benefits.Firstly,the object oriented multidimensional approach [29] is the best as it satisfies all the criteria required for the data warehouse design and it is more adaptable as the user requirements are constantly changing.Secondly,we have used UML which is easy to learn and can model all real world objects. Thirdly, star and snowflake schemas are more efficient for data warehouse design as they are easy to learn and need fewer joins [18].

In the future, we are in the process of testing the proposed framework on a case study and implementing the same using JAVA at the front-end and Oracle 10g at the back-end.

REFERENCES

[1] Inmon, W.H., Hackathorn, and R.D (1994) Using the data warehouse. Wiley-QED Publishing, Somerset, NJ, USA.

[2] June 1999,UML Modelling Language Specification. Version 1.3, Available at http://www.rational.com/uml/resources/documention / (March 2009).

[3] Booch G., Rumbaugh J., and Jacobson I.(1999) The Unified Modelling Language User Guide, Addison- Wesley Longman, p.482.

[4] Vassiliadis P. and Sellis, T.,(1999) "A Survey of Logical Models for OLAP Databases". SIGMOD Record 28(4),pp 64–69.

[5] S. Rizzi, A. Abelló, J. Lechtenbörger, J. Trujillo(2006) "Research in data warehouse modelling and design: dead or alive?" DOLAP, ACM, , pp. 3–10.

[6] A. Abelló, J. Samos, and F. Saltor (2001) "A Framework for the Classification and Description of Multidimensional Data Models" In Proceedings of the 12th International Conference on Database and Expert Systems Applications (DEXA'01).

[7] M. Blaschka, C. Sapia, G. Höfling, and B. Dinter,(1998) " Finding your way through ultidimensional data models" In Proceedings of the 9th International Conference on Database and Expert Systems Applications DEXA'98, volume 1460 of Lecture Notes in Computer science, pp 198–203, Vienna, Austria, August 1998. Springer-Verlag.

[8] Stefano Rizzi, Matteo Golfarelli. (1998) "A Methodological Framework for Data Warehouse Design". DOLAP 98 Washington DC USA.Copyright ACM 1999 l-581 13-120-8/98/l 1...$5.00.

[9] Juan Trujilio,E.Medina and S.Lujan Mora (2002) ,"A Web Oriented Approach to manage Multidimensional Models through XML Schemas and XSLT '' EDBT 2002 Workshops, LNCS 2490, pp. 29–44, 2002. Springer-Verlag Berlin Heidelberg.

[10] S.Lujan Mora and I.Song (2002),"Multidimensional Modeling with UML Package Diagrams " In Proc. of the 21st Int. Conf. on Conceptual Modeling.Lecture Notes in Computer Science pp 199-213,Finland,October 7-11,2002, . Springer-Verlag

[11] Stefano Rizzi, Matteo Golfarelli,D.Maio (1998) ”The Dimensional Fact Model:A Conceptual Model for Data Warehouses.” International Journal of Cooperative Information Systems(IJC IS),7(2-3):215-247.

[12] Lujan Mora and Juan Trujilio (2003) “A Comprehensive Method for Data Warehouse Design.”in Proceedings of 5th International Workshop on Design and Management of Data Warehose(DMDW’03), pp 1.1-1.14.

[13] Juan Trujillo and Sergio LujánMora (2004) ” Physical Modeling of Data Warehouses using UML” DOLAP’04, Washington, DC, USA.Copyright 2004 ACM 1581139772/04/0011 ...$5.00.

[14] Sergio Luján-Mora1, Panos Vassiliadis and Juan Trujillo. (2004) ” Data Mapping Diagrams for Data Warehouse Design with UML”in Proceedings of 23 rd International Conference on Conceptual Modeling (ER 04),volume 3288 of LNCS,China,Springer

[15] Lujan Mora and Juan Trujilio (2006) .”Physical Modeling of Data warehouses by using UML Component and Deployment Diagrams:Design and implementation issues.” Journal of Database Management 17(1)

[16] Rizzi,Trujilio,Abello. (2006) ” Research in Data Warehouse Modeling and Design:Dead or Alive?”. DOLAP’06, Arlington, Virginia, USA.Copyright 2006 ACM 1-59593-530-4/06/0011 ...$5.00.

[17] Deepti Mishra, Ali Yazici, Beri, Pinar Başaran. (2008) ” A Casestudy of Data Models in Data Warehousing.” 978-1-4244-2624-9/08/$25.00 ©2008 IEEE.

[18] Kamal Alaskar1 and Akhtar Shaikh. (2009)” Object Oriented Data Modeling for Data Warehousing(An Extension of UML approach to study Hajj pilgrim’s private tour as a Case Study). International Arab Journal of e-Technology, Vol. 1, No. 2.

[19] Hui Ma,Yiping Yang and Fan Zhang (2009) ”The Anti-standardized Design Research of Data Warehouse”. IEEE.

[20] Fernandez Medina et al. (2010)” Model Driven Development of Secure XML Data Warehouses: A Case Study” EDBT 2010, Lausanne, Switzerland. Copyright 2010 ACM 978-1-60558-945-9/10/0003 $10.00.

[21] Francois Pinet et. al. (2010) ”The use of UML to design agricultural data warehouses.” AgEng 2010,International Conference on Agricultural Engineering.France

[22] Jesús Pardillo and Jose-Norberto Mazón. (2011) ” Using Ontologies for the Design of Data Warehouses.” International Journal of Database Management Systems (IJDMS), Vol.3, No.2.

[23] Payal Pahwa and Shweta Taneja. (2011) ”Design of a Multidimensional model using Object Oriented Features in UML.” IARS International journal.

[24] L. Cabibbo and R.Torlone(1998) “A Logical Approach to Multidimensional Databases” in Proceedings of 6th International Conference on Extending Database Technology EDBT 98 ,Volume 1337 of LNCS,pp 183-197, Spain,Springer.

[25] J. Trujillo, M. Palomar, J. Gómez, I.-Y. Song(2001),” Designing data warehouses with OO conceptual models”, IEEE Comput. 34 (12) (2001) 66– 75.

[26] N.Tryfona,F.Busborg and J.G.Chriastiansen(1998) ,”StarER:A Conceptual Model for Data Warehouse Design”,in Proceedings of the ACM 2 nd International Workshop on Data Warehousing and OLAP , DOLAP 99,pp 3-8 .

[27] Luj´n-Mora S., Trujillo J., and Song, I.(2002), “Multidimensional modeling with UML package diagrams warehouses, ” in Proceedings of 21 st International Conference on Conceptual Modeling, ER 02 ,Volume 2503 of LNCS,pp 199-213,Finland ,Springer.

[28] Golfarelli, M., & Rizzi, S. (2001).” WanD: A CASE Tool for Data Warehouse Design”. In Demo Proceedings 17th International Conference on Data Engineering (ICDE 2001), Heidelberg, Germany, 7-9.

[29] Luj´n-Mora S., Trujillo J., and Song, I., “A UML profile for multidimensional modeling in data warehouses,” Data Knowl. Eng. 59(3) 725–769.

[30] Anjana Gosain,Suman Mann, (2010)”Object Oriented Multidimensional Model for a Data Warehouse with Operators”, International Journal of Database Theory and Application Vol. 3, No. 4.

[31] Rodolfo Villarroel, Emilio Soler, Eduardo Fernández-Medina, Juan Trujillo4,and Mario Piattini (2006),” Using UML Packages for Designing Secure Data Warehouses”, ICCSA 2006, LNCS 3982, pp. 1024 – 1034 .© Springer-Verlag Berlin Heidelberg.

[32] Payal Pahwa, Shweta Taneja and Garima Thakur(2011)” Uclean: A Requirement based Object Oriented ETL Framework”, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.2, No.4, November 2011.

[33] Sarkar, A., Choudhury, S., Chaki, N. & Bhattacharya, S, (2009) “Conceptual Level Design of Object Oriented Data Warehouse: Graph Semantic Based Model”, INFOCOMP Journal of Computer Science, pp. 60-70.

[34] Ponniah, P, (2001) Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals, pp 402.

[35] Elmasri, R. & Navathe, S.B, (2000) Fundamentals of Database Systems, Addison Weasely PubCo. ISBN 0201542633.

[36] Nazri, Mior Nasir Mior,Noah, Shahrul Azman Mohd,Hamid, Zarinah (2008),” Automatic data warehouse conceptual design “,International Symposium on Information Technology,Malaysia.

[37] Karen C. Davis Sandipto Banerjee (2007),” Teaching and Assessing a Data Warehouse Design Course”, 24th British National Conference on Databases (BNCOD'07) 0-7695-2912-7/07 $20.00 © 2007 IEEE.

[38] Mayank Sharma, Navin Rajpal and B.V.R.Reddy (2010),” Physical Data Warehouse Design using Neural Network” . International Journal of Computer Applications 1(3):86–94, February 2010. Published By Foundation of Computer Science.

[39] Abraham Silberschatz, Henry F. Korth, and Sudarshan (2002). Database System Concepts pp 445-489. 4th Edition, McGraw Hill.

[40] Marotta, A., Ruggia, R.(2002),” Data warehouse design: a schema-transformation approach”, Computer Science Society, 2002. SCCC 2002. Proceedings. 22nd International Conference of the Chilean.

Authors

Dr. (Mrs.) Rajni Jindal is working as an Associate Professor at Delhi College of Engineering(Now Delhi Technological University. She received her M.E. (Computer Technology & Applications) degree from Delhi college of Engineering. She completed her PhD (Computer Engineering) from Faculty of Technology, Delhi University in the area of Data Mining.

She possesses a work experience of around 21 years in research and academics. Her major areas of interest are Database Systems, Data Mining & Data Warehouse and Operating systems. She has authored around 40 research papers and articles for various national and international journals/conferences. She has also authored 3 books. She is a life member of professional bodies like Computer Society of India (CSI) and senior member of Institute of Electrical Engineers (IEEE), USA.

Shweta Taneja is a research scholar in Computer and Engineering Department at Delhi College of Engineering (Now Delhi Technological University. She received her M.Tech(Information Systems) degree from Netaji Subash Institute of Technology, Delhi University. Her areas of interest are Data Warehousing, Data Mining and Database Management Systems